Examination Computational Statistics

Linköpings Universitet, IDA, Statistik

Course code and name:	732A90 Computational Statistics
Date:	2019/08/26, 8–13
Assisting teacher:	Krzysztof Bartoszek
Allowed aids:	Printed books, 100 page computer document, and material in the zip file **extra_material.zip**

Grades:	A= $[18 - 20]$ points
	B= $[16 - 18)$ points
	C= $[14 - 16)$ points
	D= $[12 - 14)$ points
	E= $[10 - 12)$ points
	F= $[0 - 10)$ points

Instructions:	Provide a detailed report that includes plots, conclusions and interpretations. If you are unable to include a plot in your solution file clearly indicate the section of R code that generates it. Give motivated answers to the questions. If an answer is not motivated, the points are reduced. Provide all necessary codes in an appendix. In a number of questions you are asked to do plots. Make sure that they are informative, have correctly labelled axes, informative axes limits and are correctly described. Points may be deducted for poorly done graphs. Name your solution files as: **[your anonymous exam account]_[own file description].[format]** If you have problems with creating a pdf you may submit your solutions in text files with unambiguous references to graphics and code that are saved in separate files There are **TWO** assignments (with sub–questions) to solve. Provide a separate solution file for each assignment. Include all R code that was used to obtain your answers in your solution files. Make sure it is clear which code section corresponds to which question.

NOTE: If you fail to do a part on which subsequent question(s) depend on describe (maybe using dummy data, partial code e.t.c.) how you would do them given you had done that part. You *might* be eligible for partial points.

Assignment 1 (10p)

In statistics it is often very important to be able to estimate the probability of very extreme events, i.e. the probability that some random variable, X e.g. the time to the arrival of a very big order for a production plant, exceeds some value. These probabilities are often modelled by the so–called generalized Pareto distribution (GPD) which has the cumulative distribution function

$$F_{a,c}(h) = \left\{ \begin{array}{cc} 1 - (1 + ch/a)^{-1/c} & c \neq 0 \\ 1 - \exp(-h/a) & c = 0 \end{array} \right. ,$$

where a and c are parameters of this distribution. The generalized Pareto distribution is used to model the conditional (tail) probabilities

$$P(X \leq u_0 + h | X > u_0) \approx \beta F_{a,c}(h),$$

for some threshold value u_0 and constant of proportionality β. **For the purpose of this exam take $\beta = 1$.**

A production plant has received 17 independent orders of sizes 89, 114, 189, 76, 142, 361, 91, 178, 143, 207, 189, 268, 120, 164, 101, 178, 81. The threshold for the production plant when orders start to get big is $u_0 = 175$.

Question 1.1 (2p)

What is the density of the generalized Pareto distribution? Derive it. Implement this in a function.

Question 1.2 (1p)

Based on your GPD density in Question 1.1 write a function that calculates the log–likelihood. If you have failed to derive the density you may use `SpatialExtremes::dgpd()`. Assume that $a, c > 0$.

Question 1.3 (4p)

Use `optim()` to maximize the log–likelihood and obtain estimates of c and a. Assume that $a, c > 0$. Try the optimization out for different initial values. Can you notice anything about the estimates of c and a?

Question 1.4 (3p)

Explore the sensitivity of the estimates to the choice of u_0. Provide plots illustrating how the estimates change with the change of the threshold.

TIP: Remember that the modelling distribution is a conditional one and that there might be a reason why the order sizes were provided in the exam pdf and not as a text file that could be read directly by `R`.
TIP: Do not forget over what region `optim()` optimizes. Remember to use the constraint $a, c > 0$.

Assignment 2 (10p)

Question 2.1 (3p)

Derive the inverse cumulative distribution function method in order to draw from the generalized Pareto distribution. Implement it in an `R` function that takes a and c as parameters and then samples from the GPD.

Question 2.2 (5p)

Using your implementation in Question 2.1 use your sampler to sample a population (take 17 as your sample size) distributed according to the GPD. Take as a and c the estimated values in Question 1.3. Based on this sample estimate the probability that $P(X > 300|X > 175)$. If you failed to estimate the values of a and c, take $a = 150$ and $c = 1$. If you failed to do Question 2.1 you may use `SpatialExtremes::rgpd()`. Explore the variability of your Monte Carlo estimator, provide plots.

Question 2.3 (2p)

What is the direct estimate of $P(X > 300|X > 175)$ for the data in Question 1? Based on your exploration of the Monte Carlo estimator in Question 2.2 how confident can you be about the above estimate?